Repeated anonymised samples of administrative records:
an application to social security data in Brazil

Rigan André Campos Gonzalez (DATAPREV-Brazil)
Pedro Luis do Nascimento Silva (University of Southampton-UK)

Keywords: administrative data; microdata access; longitudinal data analysis; public use files.

1. Introduction and motivation

The Brazilian Social Security Administration (SSA) maintains huge databases with administrative records on all contributors and beneficiaries enrolled in the social security system. These databases contain a wealth of information about individuals affiliated to the General Social Security Regime (GSSR) such as sex, age, work history, wages and pension fund contributions, benefit claims and payments, etc. The records held within these databases provide a valuable source of information about labour market participation, offering in particular a longitudinal perspective that is unavailable from other sources. Statistics derived from them are especially useful when analysed in combination with those from economic and household surveys carried out to monitor the employment and coverage of social security affiliation and provision in Brazil.

Due to their sensitive and confidential nature, such large and complex databases are held in central computer facilities, under high security protection and are thus largely inaccessible for the research community. Information disseminated regularly by the SSA to the public is mostly in the form of pre-specified sets of tables, defined as cross-classifications at a high-level of aggregation. These enable monitoring of some broad indicators about the labour market and the social security system, but are far from adequate for detailed research and analysis purposes.

Anonymised samples of records can go a long way in protecting the confidentiality of individual records, while enabling the dissemination of individual anonymised microdata, as suggested by Raisinski et al. (1997). Such samples have become widely used for dissemination of census information in individual form in many countries. For example, there are now public use samples selected from every surviving census carried out in the United States of America from 1850-2000 (see http://usa.ipums.org/usa-action/faq.do). Similar samples taken since 1960 are available for many other countries from around the world (see https://international.ipums.org/international/). Another argument in favour of the dissemination of the anonymised samples is the gigantic size of the databases where they come from, which makes them inadequate sources for direct exploratory analysis and use in model fitting exercises commonly attempted by researchers.

Examples of public use samples from administrative records do not share such a long history, but are becoming more frequent. In the US, the Social Security Administration provides four such data sets covering beneficiaries of various programmes (for more detail visit their website at http://www.ssa.gov/policy/docs/microdata/index.html or see Drazga, 2008). Following debate on how to assess the development of research-ready data from state administrative sources in the areas of public assistance, public health and welfare and for use in policy and academic research, Hotz et al (1998) published a report characterizing existing state administrative databases capable of sustaining various types of research; assessing key concerns that must be addressed for administrative data to become a widely used basis for research; describing examples of where administrative data sources have been developed; identifying the strengths and weaknesses of administrative data, as compared with survey data; and making recommendations for enhancing the quality, availability and utilization of administrative data.

Also in the US, the Institute of Poverty Research of the University of Wiscosin-Madison offers a public use sample of administrative records extracted for the Child Support Demonstration Evaluation project, carried out to assess the impact of the State of Wisconsin public assistance program for low-income families with children started in 1997 (for further information see http://www.irp.wisc.edu/research/pudata/csdepud/admindata.htm).

In the UK, a similar example is the Survey of Personal Incomes, extracted by Her Majesty's Customs & Excise, and made available for research through the UK Data Archive. This survey comprises records from nearly 500 thousand tax payers in the UK for the most recent year. More information about the survey is available from the UK National Statistics website at http://www.statistics.gov.uk/STATBASE/Source.asp?vlnk=482&More=Y .

Following the lead of such examples, we developed a sampling strategy for the selection of repeated samples from the Brazilian SSA administrative records (see Gonzalez, 2005, for a detailed description). Microdata from these samples in anonymised form could be made available to vetted researchers, and perhaps even to the wider public, enabling more in-depth analysis as required. The strategy was applied to extracts from the National Database of Social Information which contain records of jobs linking employees with their formal employers for those affiliated to the GSSR, the largest social security regime in Brazil. The other large social security regime covers civil servants and military personnel in all levels of government in Brazil, and is not covered in this database, hence is outside the scope of our analysis.

The goal was to obtain stratified simple random samples of job records updated every month. The stratification was geographic (27 States) cross-classified by four broad sectors of the International Standard Industrial Classification (ISIC). The Permanent Random Number (PRN) technique – see Ohlsson (1995) was applied to perform controlled rotation of samples over time, enabling both refreshing the samples as well as keeping a short term longitudinal perspective for the samples selected at each time point.

2. Sample design and selection

The data used as a sampling frame were extracted from a repository assembled from processing the Social Security Withholding Declaration Form. These are forms which employers have to present every month containing a record for each employee they have with corresponding wages and the amount of social security contributions paid by the employer and retained from the employee for that month. The records in the sampling frame for a given month represent the employment relationships (which we call ‘jobs’ from now on) which were reported by employers for social security contribution purposes. The target population which the anonymised samples aim to represent is formed by all jobs held by workers affiliated to the General Social Security Regime (GSSR) during at least one of the months from July 2001 till June 2002.

The sampling frame was assembled in a cumulative fashion, starting with all the jobs reported in July 2001 and aggregating, one month at a time, all the new jobs reported every month, including jobs which have date of start of employment in previous months and were reported late. Jobs which ceased to exist were maintained in the sampling frame for a period of six months. After this period, they were then excluded from the sampling frame and were no longer eligible for sampling. These measures were aimed at reducing the effects of late reporting of ‘births’ and ‘deaths’ of jobs, common within the first few months following such events.

The main target for inference chosen for guiding the sample design was the estimation of the proportions of jobs in each of the following status categories (1=active, 2=new admission, 3=terminated in current month, 4=terminated in previous periods, 5=not reported). There is also interest in estimating transition rates between adjacent months, namely the proportion of job records which have status i at time t and are in status j at time $t+k$, for k=1, 2, …, 6. The estimation of such proportions would be based on the corresponding sample weighted estimators, with weights equal to the reciprocals of the unit inclusion probabilities. Additional variables of interest include the wages, sex and age of employee, duration of job, as well as some characteristics of the employer, such as location (state level only) and sector of activity.

The key domains of analysis were defined as the cross-classification of states (27 levels) and standard industrial classification of employer (four ‘sectors’, namely 1=Manufacturing, 2=Trade and distribution services, 3=Other services, and 4=Agriculture, construction and other productive activities). These four sectors were used as explicit strata in 10 states (namely Bahia, Ceará, Espírito Santo, Minas Gerais, Paraná, Pernambuco, Rio de Janeiro, Rio Grande do Sul, Santa Catarina and São Paulo). In the other 17 states, no further stratification was used, and estimation by sector of activity would proceed by using domain estimation methods, without any attempt to achieve pre-specified levels of precision. This approach yielded a total of 57 explicit strata.

Sample sizes for each of the 57 explicit strata were calculated such that proportions of jobs in each of the relevant status categories equal to or larger than 1.5% could be estimated with a maximum standard error of 0.3% at the 95% confidence level. This proportion corresponded to the smallest proportion of jobs terminated in January 2002 across the strata. The overall sample size was obtained by summing up the sizes required in each of the individual strata.

As the proportions of jobs in the various status will vary over time within each stratum, the required sample sizes in each stratum should ideally be re-calculated every time a new sample is to be selected (say, every month). However, this would require that such job status proportions be known for the whole population in each stratum every month, which may prove costly to obtain. Instead, a simplified approach was adopted where the sample sizes in all the strata are the same, and calculated such that the stated precision requirement can be satisfied all the time. This lead to using a sample size of $n = 6{,}300$ jobs in each of the 57 explicit strata, and a total sample size of 359,000 job records sampled every month.

Besides having sample sizes which provide the required precision for the monthly estimates of proportions of jobs in the various statuses, it is important to take into account the requirements for some longitudinal analysis from the sample. However, imposing similar precision requirements for longitudinal analysis lead to sample sizes which would be too large. As a compromise solution it was decided to double the sample size for the states where no stratification by sector was used. Hence the sample size adopted for each of the 17 states without detailed stratification by sector of activity was $n = 12{,}600$ job records. As a result, the total sample size each month increased to 466,200 job records. Table 1 presents the sample sizes for various domains of analysis and targets of inference.

We stress that the sample sizes we proposed to use are not guided by some optimal decisions, but offer a simple yet comfortable option that will enable some detailed data analysis to be carried out using the anonymised samples, while at the same time not being too disclosive (the overall sampling fraction is still small, i.e., less than 1.5% of the total number of job records in the database). At the same time, using fixed sample sizes across time means that no pre-processing of the database is required prior to sample selection, except for the generation of permanent random numbers for new job records added each month and the exclusion of terminated job records after their six months ‘cooling’ period.

Table 1 – Sample sizes for various domains of analysis and targets of inference

State	Sector	Sample size available for		Part of sample renewed every month
		monthly estimates	transitions 6 months apart	
Group 1 - Bahia, Ceará, Espírito Santo, Minas Gerais, Paraná, Pernambuco, Rio de Janeiro, Rio Grande do Sul, Santa Catarina and São Paulo	Manufacturing	6,300	3,150	525
	Trade and distribution services	6,300	3,150	525
	Other services	6,300	3,150	525
	Agriculture, construction and other productive activities	6,300	3,150	525
Group 2 - other states	All activities	12,600	6,300	1,050
Total across all strata		466,200	233,100	39,900

Sample selection was carried out using simple random sampling without replacement within each stratum, using the synchronised procedure described by Ohlsson, 1995, p. 165-166. Independent and identically distributed Uniform random numbers were associated with each of the records present on the sampling frame for July 2001. For subsequent months, such random numbers were held fixed for records already in the frame, and any new records would have new random numbers generated independently and from the same distribution.

The use of permanent random numbers allows for the efficient coordination of samples selected repeatedly from the same or similar populations. Their usage does not warrant for some units to stay a fixed length of time in sample, but rather, allow for some control over the expected time in sample. For the present application we proposed to rotate out 1/12 of the sample every month. This means that the expected time in sample for any job record selected is 12 months. Since there is no additional burden on those providing the information, this time can be extended if the goal is to enable for longitudinal analyses spanning longer periods of time. This would not affect the basic design and can be easily accomplished with the proposed selection mechanism.

The precise sample selection algorithm used at any given month is described in the sequence.
Step 1 – Sort the records in the updated sampling frame by stratum, and within each stratum, in ascending order of the corresponding permanent random numbers.
Step 2 – Start the process for the first selection stratum, namely h=1.
Step 3 – Calculate the rank (position) P_{hi} of each record i in stratum h according to the corresponding associated permanent random numbers. The smallest position in the stratum shall be 1 and the largest shall equal N_{th}, the total number of records in stratum h at time t. If there are any ties (which occur with probability very near zero) remove the ties by randomly assigning one of the records the smallest position (k, say), and the other to the position k+1.
Step 4 – Determine the start and end points for sample inclusion in stratum h using

$$\text{Start}_{th} = 1 + \text{mod}\left\{\left[(t-1)\frac{n_{th}}{T}\right] + 1 ; N_{th}\right\}; \tag{1}$$

$$\text{End}_{th} = \text{Start}_{th} + n_{th} - 1 \tag{2}$$

where $[a]$ denotes the integer part of a, t denotes the survey round, starting with 1 for July 2001, n_{th} is the sample size in stratum h at time t, T is the maximum number of rounds which a record is expected to be included in the sample, and mod$\{a\ ;\ \ b\}$ is the remainder of the division of a by b.

Step 5 – If $\text{Dif}_{th} = n_{th} - (N_{th} \text{ - Start}_{th} + 1) \le 0$ then include in the sample for time t the records with positions satisfying $\text{Start}_{th} \le P_{hi} \le \text{End}_{th}$. Otherwise, include in the sample for time t the records with positions satisfying $\text{Start}_{th} \le P_{hi} \le N_{th}$ or $1 \le P_{hi} \le \text{Dif}_{th}$.

Step 6 – Set $h = h + 1$ and repeat steps 3, 4 and 5 for every new stratum, until all strata have been processed.

This algorithm enables control of the sample overlap while simultaneously permitting the sampling frame to be updated through incorporation of ‘births’ and ‘deaths’, as well as any changes of strata (say an employer has its activity sector reclassified or moves from one state to another). For the application described in this paper we used T=12, namely we set the expected time in sample equal to 12 months. The sample selection algorithm proved easy to apply even with the very large databases in our example, and the intended sample rotation played its part in updating the sample.

3. Some results from the selected anonymised samples

Due to the natural 'births' and 'deaths' of jobs the observed sample renewal rates were a bit higher than the nominal rates anticipated if the population suffered no changes. We illustrate this with some observed renewal counts for selected strata presented in table 2.

One of the most important variables in the database is the status of a job at each time point. This and a series of other relevant analysis variables were recovered from the database for all records selected for the sample in any given time point. Some derived variables were also created to facilitate estimation of certain parameters.

Table 2 – Number of new job records in sample for selected strata – 2002 samples

State	Sector	Expected new sample records per month	Sample month - 2002					
			Jan	Feb	Mar	Apr	May	Jun
Rio de Janeiro	Manufacturing	525	517	674	676	628	548	567
	Trade and distribution services	525	514	691	673	630	560	581
	Other services	525	506	654	392	608	574	552
	Agriculture, construction and other productive activities	525	491	690	753	652	569	634
	All	2,100	2,028	2,709	2,494	2,518	2,251	2,334
Rio Grande do Norte	All	1,050	1,005	1,228	1,228	1,158	1,100	1,228

The selected samples were used to calculate point and standard error estimates for various target parameters. Table 3 presents some estimates of totals and proportions of jobs in each status category for April 2002, after re-weighting the sample to compensate for records with status 'not reported' or with a 'failed declaration', i.e., a declaration from which we are unable to ascertain the job status. The achieved precision was generally higher than that pre-specified for determining sample sizes, when the domains of analysis coincided with the sample selection strata or with aggregations of these.

Table 3 – Selected estimates of total and proportions of jobs by status – April 2002

Job Status	Estimated count	s.e. count	Proportion of total	s.e. proportion
New admission	1,108,620	15,728	4.20%	0.06%
Active	24,326,586	46,627	92.23%	0.08%
Terminated this month	939,685	15,326	3.56%	0.06%
Terminated previous periods	5,186,037	34,408	–	–
Total	31,560,928	–	99.99%	–

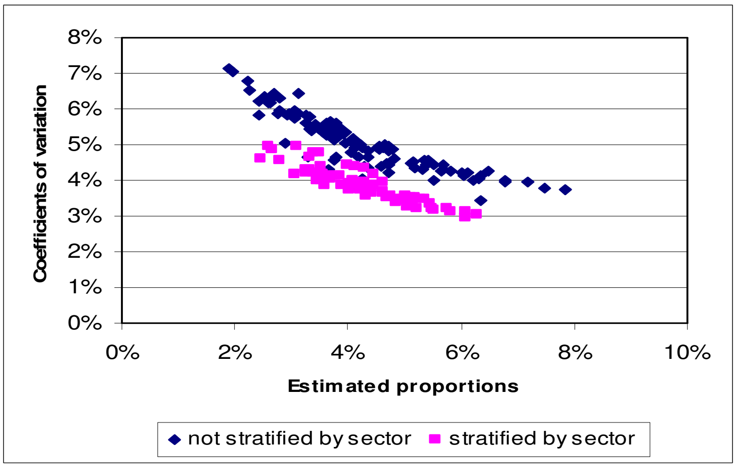
A detailed analysis of the estimates and their standard errors by the main domains of interest was carried out, but is beyond the scope of this paper (for further information see Gonzalez, 2005). To illustrate what levels of precision a user of these samples can expect to have when estimating for some of the main target parameters we provide an extract of the estimates of the number of new jobs and corresponding proportion calculated for a single state (Minas Gerais) by sector of activity (table 4). These estimates have small standard errors even for this 'rare' type of record in the database (proportions vary around 4-9% in this state for the particular month). Similar estimates for a state where no activity sector stratification was used (Goiás) are also presented in table 4 for comparison.

Table 4 – Selected estimates for the count and proportion of new jobs – April 2002

State	Activity sector	New jobs (count)	s.e. count	Proportion of new jobs	s.e. proportion
Goiás	All	37,403	1,566	6.34%	0.26%
Minas Gerais	Manufacturing	23,038	1,480	4.75%	0.30%
Minas Gerais	Distribution and trade services	26,713	1,641	5.38%	0.33%
Minas Gerais	Other services	43,573	3,167	3.91%	0.28%
Minas Gerais	Other productive activities	34,440	1,852	9.03%	0.48%
Minas Gerais	All	127,764	4,283	5.16%	0.17%

As another illustration, Graph 1 presents a scatter plot showing how the coefficients of variation (CVs) vary for the corresponding estimates of proportions of new admissions. Estimates were computed for domains defined as the cross-classification of state by activity sector, and the larger standard errors are observed for those states where the design did not stratify by activity sector.

Graph 1 – Scatter plot of CV and estimated proportions of new admissions – April 2002



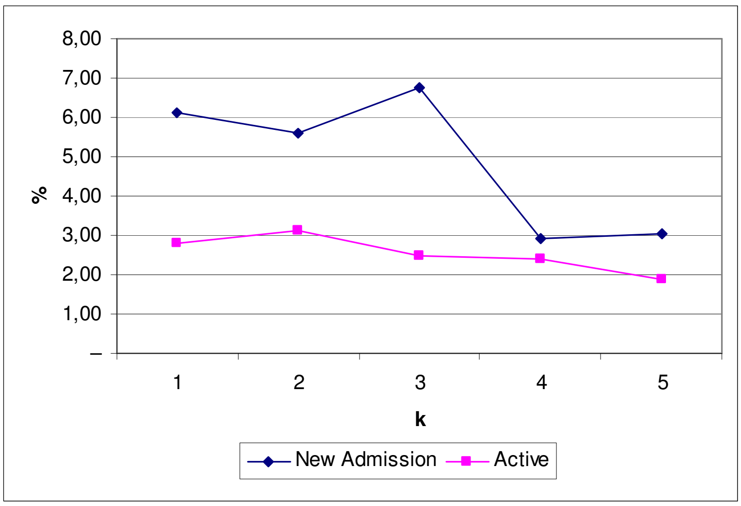
Estimating monthly variation in the target population parameters at an aggregate level is one of the intended applications that these anonymised samples might have. Hence we also examined how the samples would cope with such estimation tasks. Table 5 presents some estimates for the differences in the proportions of active jobs between adjacent months for several months. Considering a 5% significance level, the results in table 5 show that the samples would enable detection of small differences in the proportion of active jobs in adjacent months, i.e. any differences larger than 0.24% in absolute value.

Table 5 – Estimated differences in the proportions of active jobs between adjacent months

Month (t)	Proportion active (t)	s.e. proportion	Proportion active (t+1)	s.e. proportion	Difference in proportions	s.e. difference	t-test statistic
Jan-02	92.11%	0.08%	92.47%	0.08%	0.36%	0.12%	3.11
Feb-02	92.47%	0.08%	91.77%	0.09%	-0.70%	0.12%	-5.90
Mar-02	91.77%	0.09%	91.54%	0.09%	-0.24%	0.12%	-1.96
Apr-02	91.54%	0.09%	91.84%	0.09%	0.31%	0.12%	2.54
May-02	91.84%	0.09%	92.79%	0.08%	0.94%	0.12%	8.05

Another important application of the samples is the estimation of transition probabilities between job status categories (gross flows tables) for months several lags apart. We illustrate this type of analysis with the gross flows tables having January 2002 as the baseline period, and all other months in the first semester of 2002 as the ‘current’ period. Graph 2 presents the monthly evolution of the proportion of jobs terminated for jobs which existed in January 2002 (Active) and for jobs which were new admissions in January 2002 (New admissions). It shows that the probability of job termination is larger within the first three months of employment, which is in line with expectations given that the employers’ obligations are smaller during this legal probation period. Similar results were obtained for analyses having other months used as baseline (December 2001 and February 2002).

Graph 2 – Proportions of jobs terminated in month $t+k$, for jobs existing (Active) or started (New admissions) in January 2002 (k=0)



4. Conclusions and discussion

The main purpose of the administrative record system maintained within the SSA is to collect and store information that enables the administration to monitor the collection of social security contributions and to inform the concession and payment of benefits. For this reason, statistics derived from these administrative records may have limitations if one wants to understand the labour market as a whole, since many workers engaged in the 'informal economy' are not covered by the system. Nevertheless these data may still contribute to inform debate and illuminate the scene of formal employment in Brazil.

We argue that the Brazilian SSA could improve its approach for releasing statistical information by providing controlled access to such anonymised samples of microdata. This would enable satisfying analytical needs of many specialized users, while still protecting the confidentiality of individual records. Such access would substantially enhance the capacity for the study and evaluation of the impact of public policies regarding the Social Security system in Brazil. We believe similar ideas might be useful elsewhere.

The sample selection algorithm proposed worked well in our application. All the sample selection, estimation and analysis activities were carried out using a 'standard' microcomputer, demonstrating that once the samples are made available, analysts should have no difficulty in exploring the data for their own estimation and analysis activities.

The various analyses carried out with the selected samples illustrate the potential of such samples for analytical use. For cross-sectional estimates in any given month, the analyst would have a substantial sample of approximately 466,200 records, capable of delivering precise estimates even for some fine domains of interest. For longitudinal analyses, the sample size reduces for every additional period taken into account, but for samples six months apart, the analyst would still have approximately 233,100 matched records available. Our results demonstrated the analytical potential of the proposed samples, both for producing precise estimates for proportions and several other cross-sectional parameters and corresponding change over time (net change), as well as for examining short term transitions between status for individual jobs (i.e. gross flows tables).

Some important issues identified during the course of the work were left as future work. First, treatment of the problem of late job termination reporting, caused by employers not reporting job terminations for several months after the fact, is clearly an area deserving further investigation. Second is the proper assessment of the disclosure risk associated with such anonymised samples, required if the SSA decides to make them more generally available and not to restrict access to vetted analysts only. Last, but not least, is the issue of how to handle the weighting of the longitudinal samples when the records are matched or subset for analysing transitions, for example. In our initial analyses we used the naive approach of averaging the weights corresponding to the two periods involved in the analysis. However other approaches are available and their relative merits need to be investigated in more detail (see for example LAVALLÉE, 1995; FOLSON et al, 1989).

Finally, the SSA has now acknowledged this work and demonstrated interest in extracting samples from their databases, in line with some of the ideas explored in this paper. It is hoped that the approach proposed here is useful in providing at least a first step in enabling such anonymised samples to be selected and made available for analysis. This would fill in a gap in the kind of information currently available for analysts who study the formal labour market and the social security system in Brazil using administrative record sources.

References

DRAZGA, L. (2008). Uses Of Administrative Data At The U.S. Social Security Administration. Prepared for the International Seminar on the Use of Administrative Data for Economic Statistics and the Register-Based Population and Housing Census Korea National Statistical Office May 19-20, 2008, Daejeon, Korea. Available from http://www.oecd.org/dataoecd/13/53/41143137.pdf

FOLSON, R., Lavange, L. and Williams, R. L. (1989). A probability sampling perspective on panel data analysis. In Kasprzyk, D., Duncan, G. J., Kalton, G. & Singh, M. P. (eds) Panel Surveys, New York, John Wiley & Sons, p. 108-137.

GONZALEZ, R. A. C. (2005). Amostragem longitudinal em registros administrativos: uma aplicação à previdência social. Rio de Janeiro: Escola Nacional de Ciências Estatísticas, MSc. Dissertation.

HOTZ, V. J., Goerge, R., Balzekas, J. And Margolin, F. (Eds) (1998). Administrative Data For Policy-Relevant Research: Assessment Of Current Utility And Recommendations For Development. A Report of the Advisory Panel on Research Uses of Administrative Data of the Northwestern University/University of Chicago Joint Center for Poverty Research – available from http://www.econ.ucla.edu/hotz/working_papers/adm_data.pdf .

LAVALLÉE, P. Cross-sectional weighting of longitudinal surveys of individuals and households using the weight share method. **Survey Methodology** v. 21, nº 1, p. 25-32, 1995.

OHLSSON, E. Coordination of Samples using Permanent Random Numbers. In: Cox, Binder, Chinnappa, Christianson, Colledge e Kott (eds.) Business Survey Methods, New York, Wiley, p. 153-169, 1995.

RASINSKI, K. et al. Producing a public use file: a case study. Proceedings of the Survey Research Methods Section, ASA, 1997.